DOCUMENT RESUME

ED 053 142 24 TE 002 534

AUTHOR	Clark, William G.
TITLE	An Evaluation of Two Techniques of Teaching Freshman Composition. Final Report.
INSTITUTION	Air Force Academy, Colorado Springs, Colo.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
BUREAU NO	BR-5-8427
PUB DATE	Jun 68
NOTE	53p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Analysis of Variance, College Freshmen, *Composition (Literary), Correlation, *Data Analysis, *English Instruction, *Student Evaluation, *Teaching Techniques

ABSTRACT

The effectiveness of two techniques of teaching freshman composition is assessed. One technique is the use of extensive written comments on the cover sheet and pages of a theme to inform the writer of his strengths and weaknesses. The other is the use of one class period per theme for the discussion of two or three of the themes written for that assignment. The freshman course in which the research was performed was the USAF Academy Fourth Class English course. The text for the course identifies, explains, and illustrates five components of expository writing: purpose, organization, content, sentences, and diction. Effectiveness of the techniques was measured by comparing skill in using these five components on their initial and final themes. Each of three instructors was asked to teach his four classes in four different ways: teach one class using one technique; one using the other; one using both; and one using neither. Data for the experiment were grades on four of the six out-of-class themes. Four readers were employed to grade the themes according to the following criteria: (1) purpose and organization, (2) content, and (3) sentences and diction. The scores were analyzed using standard product-moment correlational analyses and analyses of covariance. No reliable evidence was found to indicate that the two techniques, used singly or in combination, were superior to instruction which offered students no guidance for improving their writing. (CK)

FINAL REPORT
Project No. 5-8427
Contract No. PO-3016-99-6

AN EVALUATION OF TWO TECHNIQUES OF TEACHING FRESHMAN COMPOSITION

June 1968

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

AN EVALUATION OF TWO TECHNIQUES OF TEACHING FRESHMAN COMPOSITION

Project No. 5-8427
Contract No. PO-3016-99-6

Lt. Colonel William G. Clark

June 1968

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such Projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

United States Air Force Academy

USAF Academy, Colorado

TABLE OF CONTENTS

ACKNOWLEDGEMENT

I wish to thank Mr. R. J. Westen, Director of Evaluation, USAF Academy, for his invaluable assistance. He designed the entire evaluation portion of the original proposal for this project; he helped in the planning and conduct of the workshop for the project theme graders, and he performed the statistical analyses of the grade data. His contribution was of the greatest importance to the success of the project.

INTRODUCTION

This project seeks to assess the effectiveness of two techniques of teaching freshman composition. One technique is the use of extensive written comments on the cover sheet and pages of a theme to inform the theme-writer of his specific strengths and weaknesses. The other technique is the use of one class period per theme for the discussion of two or three of the themes written for that theme assignment; the instructor selects themes which are representative of weaknesses or strengths demonstrated by the class as a whole and reproduces the themes on ditto or overhead projector transparency for presentation to and discussion by the class as a whole.

The extensive comment technique is the typical technique for informing composition students of their strengths and, more often, their weaknesses. When an instructor uses this technique, he tries to write comments on the theme cover sheet which will identify explicitly the strengths or weaknesses which he wants the student to recognize as a part of his writing. The instructor might find occasion to comment on all five of the components which we stress in our composition instruction; however, he is not required to comment on all of them or any specific number of them on the cover sheets of the themes he grades. The instructor uses his judgment about the comments he puts on the cover sheet and on the pages of the theme itself; on the pages of the theme he avoids duplicating the cover-sheet comments, but he tries to point out specific examples of the general strengths and weaknesses he will mention on the cover sheet as well as any others he considers worthy of comment.

The theme-discussion method requires the instructor to select from each set of themes the two or three which he feels will do the most to acquaint his students with their shortcomings and with examples of how some students have avoided or overcome those shortcomings. He always picks one excellent and one poor theme and sometimes an average one as well. In the first theme-discussion periods, he usually concentrates on clarity of purpose and relevance of the body of the theme to the purpose. Later he treats organizational features of the themes and adequacy and appropriateness of content. The instructor is free to choose whether to reproduce the selected themes on ditto or overhead projector transparencies; in one form or the other, he puts the themes before his students during the class meeting immediately prior to the submission of a new theme and discusses for that hour the matters which he feels will do the most to improve his students' writing.

The freshman course in which the research was performed was the USAF Academy Fourth Class English course, fall 1964. This course combined instruction in literature, language, and composition, with the bulk of the time spent on literature. Of the 39 class periods of that course, only 14 dealt with composition: one treated manuscript form, theme-revision, etc.; four were used for the writing

of in-class themes; two treated bibliography and documentation; two treated the basic rhetorical principles which we wanted the students to employ; and five were theme-discussion periods. The students wrote a total of ten themes, four in class and six out of class.

The short rhetoric text for this course identifies, explains, and illustrates five components of expository writing: purpose, organization, content, sentences, and diction. The two class periods devoted to rhetorical principles concentrated on these five components; the five theme-discussion periods concentrated on these five components, and the written comments which the instructors placed on theme cover-sheets and on the pages of themes concentrated on these five components.

As a result of our concentration on these five components, we felt that we could measure effectiveness of our teaching techniques by comparing student skill in using these five components on their initial themes and on their final themes. Because of the limited amount of general instruction about composition and because specific instruction was limited to written comments and theme-discussions, we felt that we could assess the effectiveness of these two techniques by asking each of three instructors to teach his four classes in four different ways: teach one class using one technique; one using the other; one using both; and one using neither.

In an effort to discriminate carefully between the two techniques and to minimize the effect of other influences, we gave all students precisely the same assignments, graded all themes according to the same criteria, and required the same revision of every theme.

The general feeling among the English Department personnel who designed the course was that the two techniques were sound and that the time devoted to composition was sufficient to provide the students with writing skill adequate to his probable requirements during an Air Force career. However, we were interested in learning if one of the techniques was better than the other or if either was weaker than we suspected. We assumed that both would prove to be of some value and that a combination of the two would be of considerable value. Feeling as we did, we recognized the need for great care in designing the project to insure that it would give us a valid evaluation of the two techniques.

METHOD

Purpose

We designed the experiment to test the relative effectiveness of two techniques of teaching English composition. The two techniques, which we have discussed more extensively in preceding paragraphs, were used independently and in combination in the various classes and are designated for purpose of design exposition as: (1) Combined Method (this method used both the theme-discussion technique and the extensive written comment technique); (2) Theme-Discussion Method (this method uses the technique of discussing two or three student themes in class); (3) Extensive Comment Method (this method uses the technique of extensive, written instructor comments about the strengths and weaknesses of each theme); (4) Control Method (cadet themes in this group receive no theme-discussion and, at most, brief, general written comments on their themes, e.g., "Good job!" "Be more careful next time." "Generally Satisfactory.").

Students

The experiment involved twelve sections of classes of the Air Force Academy's regular Fourth Class (freshman) English course, with three sections being taught by each of the four methods. Students were assigned to the sections on an essentially random basis. (Actually assignment was based on scores of a test which in past years had proven uncorrelated with English composition grades.)

Although each of the sections initially enrolled 11 to 15 students, usable data were available on only 10 students per section. Part of the reduction in sample size was due to resignation or discharge of a few cadets. The remaining losses occurred through absence from class (and consequent unavailability of data or lack of necessary student participation) of a few cadets, and participation in extra instruction or tutoring of a few others. Such participation, or lack of it, "contaminated" in an unknown fashion the teaching methods of the experiment, and the cadets involved were excluded from the sample. For ease in analysis of the data, each experimental section was reduced, by random selection of cases to be discarded, to the N of the section having the smallest number of usable cases: 10. As a result, the final experimental sample contained 120 students.

Instructors

Three instructors handled all instruction of the experimental sections. Each instructor taught four of the sections, each section by a different method. The three instructors were experienced and had all attended at least two summer workshops during which the techniques being evaluated were explained and discussed, and all

three had used both techniques in at least one year of teaching Freshman English at the Academy. The instructors received no additional formal training except for two meetings in which the project was explained and their particular responsibilities discussed. We asked these three instructors to participate in the project because we were, through personal acquaintance, assured of their competence and conscientiousness.

Course Content

Of the 39 class meetings in the course, 18 were discussions of literature, five discussions of language, 5 theme-discussion periods, 4 in-class theme periods, 3 discussions of bibliography, documentation, manuscript form, etc., 2 discussions of rhetorical principles, 1 a mid-term examination, and 1 an instructor option period during which the instructor discussed some portion of the literature treated in the course.

Each section had the same assignment for every given class meeting (except where the requirements of the experiment ruled out the theme-discussion periods); each student wrote four themes in class and six themes out of class, and the assignments for all ten of these themes were the same for all students.

Because of this uniformity of assignment and class activity and the emphasis on literature (the 5 theme-discussion periods became literature periods in those sections which were not using the theme-discussion method), we felt that the factors influencing student writing were extremely limited, coming primarily from the theme-discussion periods, the instructor comments on themes, or a combination of the two. As a result, we assumed that the experiment would provide a valid evaluation of our two techniques of teaching composition.

Data

Data for the experiment were grades on four of the six out-of-class themes; we used the first, second, fifth and sixth themes. The first two themes, submitted at lessons 5 and 14 respectively, provided measures of proficiency before training, and the last two themes, submitted at lessons 33 and 37 respectively, provided measures of proficiency after instruction. The period of differential instruction for the four groups spanned the 19 lessons between the second and fifth themes.

All of the themes were essentially expository, although #2 was potentially argumentative. Each assignment required the student to write about some aspect of the literature he was studying, and most of the assignments provided some limiting of the subject and some general structuring of the topic. Most of the assignments offered advice about narrowing, about developing ideas, and

about aspects of the subject which the student might profitably examine and discuss. The purpose of providing this guidance was to give the student maximum assistance with the difficult aspects of the task of composition during his initial writing experiences. At the end of the semester, the final assignment required the student to make his own decisions about limiting, structuring, and developing his subject.

Theme #1 was a 500 word theme on the subject of Holden Caulfield's movement "from a state of innocence into the world of experience."

Theme #2 was a 700 word research theme on the subject of "the effectiveness of the ending of Huck Finn." Each student had a text which included the novel and a collection of critical essays about the novel, and had read several of the relevant essays in preparation for class discussion of the novel.

Theme #5 was a 500 word theme in which the student chose one of the social evils which Steinbeck attacked in The Grapes of Wrath and explained how Steinbeck "conveys the full force of the social evil to the reader."

Theme #6 was a 500 word theme in which the student explained Faulkner's view that the white man in the South lives under an inherited curse and discussed the validity of that view as it operated in Light in August. (This is the assignment mentioned above which put the student entirely on his own so far as limiting, organizing, and developing his subject is concerned.)

The instructors involved in this project graded their themes in their customary fashion for actual use in the course, but these grades did not become a part of the data for the project. Instead, the themes were reproduced in several copies exactly as the students had written them, (except that all were now typed), but without any instructor marks, comments, or grades. These copies of the themes were identified only by a code number which permitted identification of the student authors and the theme number by the experimenters but which did not identify to anyone else either the authors or the theme numbers. At this point the themes were ready for distribution to the project's four readers.

I selected the four readers with the assistance of the Assistant Superintendent, Colorado School District 20. He made available to me his file of qualified high school teachers who were not then teaching full-time but were interested in teaching on a substitute basis. With his help, I selected five likely candidates and interviewed them. One was not interested, but the other four were, and they agreed to attend a two-week workshop in theme-grading and then perform the grading of 420 themes in the manner and according to the criteria I would acquaint them with during the workshop.

I conducted the first week of the workshop during mid-December, 1965, and the second week early in January, 1966. One of the four graders missed one day of the workshop; the other three missed no meetings.

The workshop undertook to acquaint the graders with three separate grading categories, one at a time. The first was the category "purpose and organization." The second was "content." The third was "sentences and diction."

I provided the graders with copies of the rhetoric text the students had used and with the definitions of the categories which the instructors used. We discussed the rhetoric section, examining carefully each category and the definition of each category, and considering how best to apply this information to the grading of a theme. Then the graders began examining actual themes and assigning grades for the category under consideration. These themes were themes written by members of the fourth class other than those involved in this project. At the beginning of the workshop, we used themes which had been written at different times and on different topics than those used in this project, but as the workshop progressed we began to use themes written on the four topics which had provided the themes for the project. (At no time did we actually use a theme which was a part of the data for the project.) I participated in their discussions of the results for the first few themes, but soon was able to let them do all of the discussing themselves. All four were competent and conscientious and questioned one another searchingly whenever there were significant differences in their evaluations of a particular theme's quality. In addition to the themes they graded in the workshop, they took themes home with them and graded them at night in preparation for the next day's workshop meeting.

We followed this same procedure for each of the three categories separately, and then began grading some themes in all three categories to get an overall grade. Gradually the graders came close together and stayed consistently so except for an occasional deviation. Always they gave the theme on which a discrepant grade [illegible] a thorough analysis and discussion in their effort to find the constituent of that theme which had caused one of them to react differently from the others.

At the end of the second week of the workshop, they and I were convinced that they were in essential agreement about what the three categories involved and what constituted excellence, adequacy, etc. in each of the three categories. At this point, we explained the procedure for grading the themes which the students in the project had written; we then gave each grader copies of all 480 themes and asked him to grade them and return them within approximately three months.

We asked the graders to mark each theme with three separate scores: the first score (Part 1) was an evaluation of purpose and organization, the second (Part 2) an evaluation of content, and the third (Part 3) an evaluation of sentences and diction. The total score (T) was the sum of the other three scores.

In our instructions, we asked the graders to use two special procedures which we felt would materially improve the prospects for reliable grading. The first procedure called for three separate readings of the entire set of 480 themes. We asked that they first go through the entire 480 themes to grade them for purpose and organization, then go through them all again to grade them for content, then again for sentences and diction. They were to record each of the three grades on the back of the themes where they would not see them at the time they were deciding on a new grade. We hoped that this procedure would assist the graders in concentrating on one category at a time and simplify their task of eliminating from their consideration the other components of composition.

The second special procedure which we recommended to the graders was the sorting of the themes into five equal-sized merit groups as they read them, rather than assigning a numerical grade to each one as they read it. This system required the graders, in effect, to decide whether the quality of the purpose and organization of a given theme placed it in the top 20% of the 480 themes, the second 20%, the third 20%, the fourth 20%, or the bottom 20%. When the themes were in these five, relatively equal piles, the graders assigned the grades appropriate to each pile. After assigning the grade for purpose and organization, the graders were asked to shuffle the 480 themes together in random order and begin grading them for content, using the same system of distributing them into five separate quality groups. They were asked to use the same system for evaluating sentences and diction. We hoped that this procedure would reduce the "halo effect" from one category to another, thereby increasing the validity of the category by category grading. This rank-ordering by fifths also insured use of the entire grading spectrum and eliminated the possibility that the instructors might assign quite different grades to a theme whose relative quality they actually agreed on.

Since the distribution of scores was specified in advance, the opportunity for inter-reader agreement was enhanced. It is obvious that if one grader gave predominantly high marks and another predominantly low marks, chances for substantial agreement between the two would be small. However, we did not insist that the themes be placed in five exactly equal piles; we authorized the graders to depart from this suggested distribution of 20% per pile whenever they believed it necessary.

After the graders finished their grading and returned the graded themes, personnel from the Academy Evaluation Office sorted the themes according to author and theme number and collated the

scores, with each grader's scores separately identified. Each theme then had 12 part scores and four total scores, 16 scores in all, four from each grader.

Statistical Analyses

The experimental design called for a series of steps in analyzing the data, with each step helping determine the nature of subsequent analysis procedure. The anticipated possible outcomes of each step were laid out in a tabular form with the decisions considered appropriate to each outcome specified in advance.

Standard product-moment correlational analyses were used for the first three steps. The first question to be answered was to what extent the graders agreed in terms of the three part-scores and the total score over the entire set of 480 themes. Had there been no statistically significant evidence of agreement among graders, further analysis would have been meaningless, and the project would have terminated at this point. On the basis of these same correlations, decisions were to be made as to whether or not the grading of any one of the four graders differed markedly from the other three. Had one grader deviated significantly from the other three, that grader's scores would have been discarded.

Again, using product-moment correlations between scores as a guide, decisions were to be made concerning independence of scores and whether or not part-scores should be combined or treated independently. Another question to be answered was whether two separate sets of analyses of covariance should be run, the first set using theme #1 scores as independent variables and theme #5 scores as dependent variables, and the second set using theme #2 scores as independent variables and theme #6 scores as dependent variables, or whether analogous scores from themes #1 and #2 should be combined into a single pre-instruction independent variable to be analyzed against an analogous combination of scores of themes #5 and #6, serving as the post-instruction measure.

Following these preliminary analyses and decisions, the experimental design called for analyses of covariance of whatever part- or total scores has proved defensible as experimentally independent, reliable measures of proficiency in theme-writing, with pre-instruction scores used to adjust post-instruction scores for any initial differences in proficiency level between the four groups of students taught by the different instructional methods.

Hypotheses

It was originally hypothesized that the different methods of instruction would result in differences in proficiency between the different groups of students taught by the different methods,

and that the differences in proficiency would be apparent in one or more of the three categories which the graders had examined and which had been emphasized for the students by both of the instructional techniques under consideration. However, the direction of any possible differences and the superiority of a particular technique were not hypothesized.

In designing the project, we assumed implicitly that graders given a short training course could grade each of three specified categories of composition with a statistically significant degree of agreement among them. Further, we assumed that there would be statistically significant independence between the three part-scores which resulted from the grading of the three categories. These two assumptions, considered simultaneously, imply that the relationship between graders' scores for a single category on one theme should be closer than the relationship among the scores a single reader assigns to the three categories of that theme.

RESULTS

Statistical analyses of the grades which the four graders assigned to the 480 themes fail to show any significant difference between sections which had received the "benefit" of instruction which employed either or both of the techniques which we were examining and the sections which received no special help from their instructors. That is, the students who wrote ten themes without the benefit of theme-discussion periods which would show them what strengths and weaknesses were present in their writing and without the benefit of detailed instructor comments about the strengths and weaknesses of each of their themes made substantially the same improvement in their writing as did those students who participated in a theme-discussion period before writing each out-of-class theme or those students whose instructors wrote extensive comments about the strengths and weaknesses of each of the themes they wrote or those students who participated in theme discussion periods and also received extensive written comments on their themes. All four of the groups (the three method groups and the control group) made about the same progress.

While progress appears to be significant for all four groups, this finding is open to some question because it depends on subjective judgments of the relative difficulty of the four theme assignments. We assume, for example, that theme #6 was quite difficult for all students because grades on that theme were consistently lower across the board than they were on theme #5. On the other hand, the grades on theme #5 were appreciably better than those on themes #1 and #2, and we attribute this difference to improved writing skill, believing that themes #1, #2, and #5 were of about the same difficulty. Because of our uncertainty about the relative difficulty of the assignments, we cannot make any confident assertions about the extent of the students' writing improvement. We believe that they all improved in about the same degree, but we don't know what that degree is.

Since no experiment can prove a null hypothesis, we cannot prove that the two instructional techniques are of no value in teaching the skills of English composition. We can only show that, in the kind of course we have described, with the kinds of students used in the experiment and with the criteria for assessing the effectiveness of writing that we used, we found no reliable evidence that the two techniques, used singly or in combination, were superior to instruction which offered students no guidance for improving their writing.

The Theme-Grading

The results of the four graders' grading of the 480 themes were not as reliable as we had hoped they would be after seeing

the four graders at work in the workshop. Analysis of the part scores which the graders assigned to the themes showed that there was a higher correlation between a reader's score for Part 1 and his score for Part 2 of a particular theme than there was between his score for Part 1 and another grader's score for Part 1 of that theme. The average inter-reader correlation of part scores was .304; this correlation, though statistically significant, is low.

The inter-reader correlation of total scores was somewhat higher, (.360), but still lower than we had hoped. By combining the four total scores for each theme, we achieved an estimated reliability coefficient of .620.

The Analyses

The analyses of the grade data were substantially as planned. Because our assumptions about the general similarity of themes #1 and #5 and of themes #2 and #6 proved in error, we did not make the planned analyses of co-variance of the two sets. Instead we combined the scores for themes #1 and #2 to produce a single pre-instruction score and the scores for themes #5 and #6 to produce a single post-instruction score. (This pooling further increased the reliability of the measurements.) The analyses of variance and co-variance of these pre-instruction and post-instruction scores plus several additional analyses led to the afore-mentioned results as the only possible conclusion which the data would support.

Appendix B contains a detailed discussion of the analyses.

DISCUSSION

Qualifications

The results of this research are sufficiently unexpected and disturbing to call for a careful look at all aspects of the project which might have in some way inadvertently influenced the project. As we re-examine the project as a whole, three aspects stand out as possible sources of uncontrolled influence: the three instructors, the theme assignments, and student motivation and effort.

The instructors were both competent and conscientious, but they had to rely upon their own interpretations of how to conduct theme-discussion periods and of how to comment on cadet themes. They had had some instruction in these matters at two or more summer workshops for all instructors involved in teaching fourth class cadets, but they did not get any special, supervised instruction in commenting on sample themes or conducting practice theme-discussion periods. Probably an opportunity to do both the commenting and the conducting of theme-discussion periods in the presence of and with the suggestions of the other two instructors would have insured that each instructor was aiming at precisely the same goals and using the same methods as the other two and that all three were performing in consonance with the exact requirements of the project. The similarity of results from each instructor's sections suggests, however, that such practice would probably not have made a significant difference in the results.

The second area where more careful advance preparation might have been influential is the set of theme assignments. Certain goals of the course dictated the nature of the theme assignments, and the final products were in some ways out of keeping with the requirements of the project. Theme #2, for example, was a research paper, which none of the others was; it was also somewhat longer than the other three. In addition, the instructions in that theme #2 assignment were considerably more detailed than in any of the other three. Conceivably, these factors may have in some way done more to shape the quality of the themes written for this assignment than did the comments the instructors had written on theme #1 or the discussions of theme #1 which the instructors had led.

The assignment for theme #6 differed from the other three in a significant way. Where all of the earlier assignments had provided detailed guidance to the cadet about how to limit, arrange and develop his theme topic, the assignment for theme #6 was deliberately general (as well as being based on a difficult novel and requiring student understanding of a rather complex idea). The decision to make the topic general was a result of the course planners' desire to find out if the students had learned from

our guidance how to limit, arrange, and develop their topics or if the second semester would need to put more emphasis on those features of composition. In the judgment of the instructors, most of the #6 themes written in the course as a whole proved to be weaker than the earlier themes in the course in all three of these areas; evidently, the students had not learned to do these things for themselves. As a result, the quality of the #6 themes which the students in this project wrote may conceivably have been determined more by the generality (and difficulty) of the assignment than by the comments and theme-discussions which they had received during the course of the semester.

The area of student motivation and effort is one which we could have done nothing to control. The mood of a student when he sits down to write a theme, the amount of time he has reserved from his busy schedule for the writing of the theme, the successes or failures of his various ventures during the preceding hours or days, and his knowledge of how much effort he needs to expend on the theme to achieve whatever grade he will be content with in the course--all of these elements affect in some indeterminate way the quality of each theme the student writes. The variation in the grades individual students received on their four themes was so great and so unexpected that we feel certain that this uncontrolled area of motivation and effort exerted considerable influence on the final outcome of the experiment. Unfortunately, we can think of no way to control this factor in future experiments, nor of any way to adjust for it in making analyses of the grades the students earn.

It is, of course, too late to make changes that will alter the results of the project, but because of the nature of the results, it seems desirable that other investigators make a further examination of the influence of these or similar techniques on student writing ability, and such an examination should profit from our experience by insuring stricter control of the research. The instructors teaching the classes should have a workshop of at least one week, during which time they practice writing comments on practice themes and discuss their comments with one another; they should also hold theme-discussion periods with actual classes, taking turns leading discussions and observing others leading them. They should follow up each class with a detailed discussion of the leader's conduct. In effect, this procedure would give each instructor the benefit of the others' views and experience and provide all of them with the same frame of reference.

The director of the research should be careful to design all of the theme assignments to make them as nearly equal in all aspects as he possibly can. Such equality will reduce the likelihood that the quality of the themes is the result of unanticipated and uncontrollable factors rather than of the techniques under examination.

The director of research must recognize the problem posed by the influence of student motivation and effort and give careful consideration to every possible solution to the problem. It may well be, however, that he will find it necessary to leave the problem unsolved and hope that the influence does not cause an unrecognized but significant distortion of his results.

Appendix C contains a discussion of the correlational analyses of the data which the graders produced and additional discussion of student motivation and effort.

CONCLUSIONS

This project sought to learn if two specific techniques of teaching freshman composition were productive. One of the techniques called for the instructor to write extensive, constructive comments on the cover sheets and the pages of his students' themes informing the students of the strengths and weaknesses of each of their themes. The other technique called for the instructor to conduct a theme-discussion period on the lesson before a theme was due. For these periods, the instructor would select themes from the previous set which his students had written, reproduce these themes and place them before his students for detailed discussion of strengths and weaknesses.

This examination of the effectiveness of these two techniques of teaching freshman composition found neither of them to be helpful. The students who received the benefit of the two techniques did not produce significantly better writing than the students in the control group, who received the benefit of neither of the techniques.

SUMMARY

The purpose of this project was to examine the effectiveness of two methods of teaching freshman composition. One method is the use of extensive instructor comments on theme cover sheets and on the pages of the themes to inform students of the strengths and weaknesses in their writing. The other is class discussion of representative themes taken from each set that the class produces; here too, the emphasis is on the strengths and weaknesses in the writing. Both techniques aim at persuading the student to adopt the strengths and eschew the weaknesses.

To assess the effectiveness of the techniques, we asked three USAF Academy instructors who were experienced in teaching freshman composition to Academy cadets to conduct their classes by four different methods. Each instructor taught one class using only the extensive comment technique; he taught one class using only the theme-discussion technique; he taught one class using both techniques, and he taught one class using neither technique. Except for these techniques, the classes had only two class-periods devoted to principles of composition, two devoted to bibliography and documentation, one to manuscript form, theme-revision, etc., and four to writing in-class themes. The students wrote six out-of-class themes. Each theme assignment was the same for all students in the course. For students who received theme-discussion, there were five class periods devoted to that technique. All students spent 25 periods on literature assignments, and those who did not participate in theme-discussions spent an additional five hours on literature.

At the end of the semester, after the instructors had assigned their grades, the themes which the project students had written were separated from those of the rest of the students in the course, and out-of-class themes #1, #2, #5, and #6 were typed on mats and reproduced in several copies, with all of the students' singularities retained as faithfully as the typist could retain them.

Each of the twelve project classes had started the semester with about fifteen students. By the end of the semester, all twelve classes were, for the purposes of this project reduced to the number of students in the class having the smallest number of "uncontaminated" students (i.e., students who had missed no relevant classes and had received no special instruction about composition). This class size was ten. This reduction gave us a total of 480 themes, four from each of 120 students.

From a list of substitute English teachers provided by a local Colorado school district, we selected four experienced teachers to serve as theme-graders. These four graded all 480 themes, using criteria we provided them; they did not know the sequence in which themes had been written nor by what method the writer of any theme

had been instructed. We gave these four graders a two-week workshop in grading, introducing them to the particular principles which had governed the instruction in each of the two techniques and having them convert these principles into the criteria by which they would grade the 480 themes. In the workshop, they spent most of the two weeks grading sample themes and comparing their grades in an effort to achieve reliable results. At the conclusion of the workshop the graders took the project themes with them, graded them according to our instructions and returned the graded themes in slightly less than four months.

The graders' grades on the 480 themes proved to be sufficiently reliable to permit us to draw definite conclusions about the efficacy of the two techniques of teaching freshman composition in the particular situation which we set up. (The reliability of the grading was not as high as we had hoped. The grades would not, for example, have permitted us to make assertions about what aspects of writing the techniques had influenced most if the experiment had shown either or both of the techniques to be of some special benefit.)

The results of the experiment show that the use of the two techniques, either singly or in combination, did not produce any more improvement in student writing than did the control group method of requiring the students to write the ten themes but giving the students no more assistance in improving their writing than marking mechanical errors and placing a grade on each theme.

The conclusion that these two techniques are of no special benefit in the teaching of freshman composition cannot, of course, be applied to all situations because no experiment can establish a null hypothesis. The result is worth noting, however, and certainly suggests that additional research in this area would be advisable. If these techniques, which logic, sentiment, and the subjective experience of thousands of composition teachers insist are valuable, are in fact worthless, we need a careful examination of the entire methodology of composition instruction. If the methods we have relied on for years are not fruitful, then we need to learn what methods are.

BIBLIOGRAPHY

Braddock, Richard; Lloyd-Jones, Richard; Schoer, Lowell, Research in Written Composition. Champaign, Illinois: National Council of Teachers of English, 1963.

Godshalk, Fred I.; Swineford, Frances; Coffman, William E., "The Measurement of Writing Ability." Research Monograph, Number 6. New York, New York: College Entrance Examination Board, 1966. 84 pp.

Lindquist, E. F., Design and Analysis of Experiments in Psychology and Education. Boston, Mass.: Houghton Mifflin Co., 1953.

Appendix A
Proposal S-403-65
Project S-428

Project Title:	An Evaluation of Two Techniques of Teaching Freshman Composition
Submitted by:	The United States Air Force Academy, USAF Academy, Colorado
Initiated by:	Major William G. Clark, English Department, USAF Academy. Area 303, 472-4589 or 472-3930
Fiscal Officer:	Accounting and Finance Officer, USAF Academy, Colorado Telephone: 472-3160
Date Transmitted:	8 February 1965 (Revised: 19 August 1965)

1. Abstract.

a. Objectives:

This project is designed to assess the effectiveness of two techniques of teaching composition. One technique is the familiar technique of writing extensive comments on the pages of themes and on the themes' cover sheets to inform the student of his specific strengths and weaknesses. The other technique is the use of one class period per theme for the discussion of two or three of the themes written for that theme assignment. The instructor selects themes which are representative of weaknesses or strengths demonstrated by the class as a whole and reproduces them on ditto or overhead projector transparency for presentation to the class as a whole. He and his class analyze the themes thoroughly, examining with special care those components which the course rhetoric identifies as most significant to effective writing.

b. Procedures:

Four groups of approximately 30 students each are participating in the project. With group one, instructors use both techniques. Instructors of group two use only the thorough comment technique, and instructors of group three use only the theme discussion technique. Instructors of group four use neither technique. The bulk of instruction aside from that described above is in the areas of literature or language, and all students cover the same material in these areas. The students write a total of 10 themes, and for any given theme, all students write on the same topic.

The first two themes and the last two themes of all students participating in the project will be reproduced on ditto, unidentified as to writer, theme number, or experimental group, and given to four graders. Each grader will grade all of the 480 themes, and their grades will be analyzed by USAF Academy statisticians to determine what effect, if any, the two techniques have on student writing skill when used either separately or in combination.

2. Problem.

Composition instruction in the USAF Academy Fourth Class (Freshman) English course is based on the assumption that competent expository writing consists of several essential components which can be identified, defined, and explained. We further assume that students can learn to assemble these components into an effective theme, and that they learn best when they know what the components are, practice using them, and have their practice evaluated.

We have identified and defined five components: Purpose, Organization, Content, Sentences, and Diction. We have written a brief rhetoric organized around these five components. We have developed an evaluation system which requires the instructor to consider these five components in arriving at a theme grade. We have designed a theme cover sheet which shows the student what weights we attach to the five components and also how, on a particular theme, the instructor evaluated the components in arriving at the grade he assigned the theme. Each instructor sets aside one class period for discussing each set of themes. The instructor brings to class mimeographed copies of two or three themes written by students in that class. He usually picks one good and one bad theme and perhaps one mediocre one. He spends the period discussing the strengths and weaknesses of these themes in terms of the component areas. Within the context of our method, there are two instructional techniques which we wish to evaluate: 1) the theme discussion technique mentioned above; 2) the use of extensive, specific instructor comments on the theme cover sheet and on the pages of the theme itself.

In our Fourth Class English course, we devote most of our time to literature and relatively little to any kind of formal instruction in composition. Of the 39 hours of our first semester, one assignment covers our brief rhetoric; one covers manuscript form, grammar, punctuation, spelling, and theme revision; one covers bibliography and documentation; six assignments require the writing of themes; six cover a language text; 22 cover novels; one is an instructor option, and one is review for a mid-term exam. The actual class activities during these 39 periods consist of 18 periods discussing literature, five periods discussing language, five theme-discussion periods, four periods writing in-class themes, two discussing rhetoric, two discussing bibliography and documentation, one taking a mid-term exam, one discussing manuscript form, theme revision, etc., and one an instructor option period.

As a result of this emphasis on literature, the number of factors influencing student writing skill is small. Most of the influence should come from either the theme-discussion periods or the instructor comments or a combination of the two. Since the other factors operate on all students about equally, we hope that by giving different groups different exposures to the influences of theme discussion and theme comments, to determine what values these techniques have in teaching a student to write.

Both of these techniques demand much time and effort from the instructor. If they produce significant improvement in students' writing ability, the time and effort are well spent. If they do not produce significant improvement, we are wasting many hours of instructor and student time. Because of our standardization of the Fourth Class English curriculum and instruction,

we are in a unique position to examine the value of these techniques. All of the students have the same assignment for each lesson; all of them write the same number of themes on the same topics and submit them at the same lessons. All instructors cover the same subject matter in class (though their methods and emphases are their own), grade themes according to the same criteria, and use the techniques of extensive comment and theme discussion periods.

We are keenly interested in learning how valuable these techniques actually are. Since they demand so much from the instructor, we strive to keep the instructor load low to permit him to do a thorough job of employing these techniques. Should the study show that they are truly valuable, we will continue our efforts to maintain a low student load. Should the study show that the techniques make no real difference in the students' writing skill, we would drop them and either seek new techniques or change the instructor's load.

3. Related Literature.

One study which relates in part to the objectives of this project is that by Earl W. Buxton: "An Experiment to Test the Effects of Writing Frequency and Guided Practice upon Students' Skill in Written Expression." This study is an unpublished dissertation, Stanford University, 1958. It is summarized and analyzed in the NCTE publication Research in Written Composition, by Richard Braddock and others, 1963.

The papers of one of the two writing groups in this study (the Revision group) received internal comments and marks as well as one paragraph each of general evaluation; the papers of the other group (the Writing group) received internal marks and the paragraph of general evaluation. The Writing group's papers received no grades, were not discussed in class, and required no revision. The Revision group's papers received grades, and one period per theme assignment was used in the following way:

> The papers of the students in the Revision group were returned at the beginning of a class period. The general strengths and weaknesses of the essays were pointed out at that time, and excerpts exemplifying certain good features were read to the class to elicit comments on how the effectiveness was achieved. Then the students were required to correct the errors indicated on their papers while the reader went from student to student giving assistance where it was needed. No more than one 50-minute class period was devoted to these procedures, the average time having been 35 minutes for each assignment. (Research in Written Composition, P. 61)

The Buxton study did not attempt to isolate influences to determine the effect of particular ones; instead, it gave one group the benefit of a careful discussion of the theme assignment, suggestions for planning and organizing, grades, internal and external comments, oral discussion of general strengths and weaknesses, oral reading and group discussion of good features,

and the requirement for some sort of supervised revision. It gave the other group nothing but the paragraph of external comment.

The project proposed herein seeks to discriminate carefully between the two techniques under consideration and to minimize the effect of other influences. To this latter end, we give all students precisely the same assignments, grade all themes on the same criteria, and require exactly the same revision of every theme.

While the Buxton study concludes that the Revision group profited significantly more than the Writing group, it cannot identify the contribution which any of the several possibly significant influences may have made to the improvement of the Revision group. If we can establish that the students in the groups taught by either or both of the techniques under evaluation have improved their writing skill significantly, we will be able to say with some confidence that one or the other or both of the techniques were responsible and perhaps that one technique is of greater value than the other.

4. Objectives.

This project seeks to determine if the following teaching methods significantly influence student writing ability:

a) The use of extensive, specific instructor comments on the theme cover sheet and on the pages of the theme.

This technique is presumable the obvious choice of the conscientious instructor, and common sense suggests that it does significantly improve the students' writing ability. In this project, the instructor comments on the components in the appropriate sections of the theme cover sheets,* attempting to identify for the student the important strengths and weaknesses his theme contains in each component area. In addition, the instructor writes comments at appropriate places on the pages of the theme; these comments are not necessarily identified explicitly with a particular component, but the connection is usually unmistakable. Both cover sheet and in-theme comments are as frequent, extensive, and detailed as the quality of the theme requires and the instructor's schedule permits.

b) The use of one 50-minute class period at the time of the return of each set of out-of-class themes for the discussion of typical strengths and weaknesses found in that set of themes.

This technique is not, so far as we can tell, widely used, but it is fairly simple, the only significant drawback being the time required to prepare copies of the themes to hand out to students.

In using this technique, the instructor selects two or three themes which he feels typify the strengths and weaknesses of his students on that particular theme topic. He reproduces these themes on ditto or on transparencies for use with an overhead projector. In either case, he places the

*A copy of the theme cover sheet is attached.

themes before his class and goes through them slowly and carefully, pointing out what features contribute to or detract from the quality of each component. In the first such class meeting, the instructor does most of the work because the students are not yet sure exactly what the instructor thinks is important, and the few comments students make are usually about punctuation, spelling or grammar.

During the subsequent theme discussion periods, however, the instructor relies on his students for much of the discussion of the themes, expecting more insight into the nature of effective writing with each successive period. For example, he asks his students to use the techniques he uses when actually grading a theme: identify the central idea and then measure the relevance of the key points in the body by comparing them to that central idea; assess the probable adequacy and interest of the illustrative material for the average, mildly disinterested reader; consider the amount of assistance which the transitions give the average reader, etc.

Common sense suggests that these two methods must surely improve students' skill as writers. Common sense also suggests that the world is flat. We would profit from more reliable information about the value of these techniques.

5. Procedures.

The project involves a total of twelve classes which began the semester with approximately fourteen students each. The twelve classes were divided into four groups of three classes each, and three instructors taught one class each from each of the four groups. All twelve sections covered exactly the same material; all wrote the same number of themes, wrote on the same subjects and submitted their themes at the same lessons.

The differences in the groups' instruction were limited to the instructors' written comments on the themes and to in-class discussions of the themes:

1) Group One: The instructor wrote on the theme cover sheet extensive, specific comments about each of the areas he had evaluated in arriving at his grade. He commented on strengths as well as weaknesses. He also commented extensively at appropriate points in the body of the theme; these internal comments further clarified the criticisms and commendations on the cover sheet.

The instructor set aside five lessons for the discussion of five sets of themes. All five of these themes were written out of class. He mimeographed (or prepared transparencies of) two or three of the themes he was returning and presented these to the students. One theme was good, and one poor. The third, if used, was average. He spent one period discussing the strengths and weaknesses he found in the themes or let the students comment on and perhaps evaluate the areas they knew the instructor examined in grading their themes.

2) Group Two: The instructor wrote extensive comments on the cover sheet and in the body of the theme. He did not conduct any theme-discussion periods.

3) Group Three: The instructor devoted five periods to the in-class discussion of five sets of themes. He limited his cover sheet comments to a single, general sentence or phrase for each of the areas he evaluated, and he limited his comments within the theme to symbols, abbreviations, or single words (?, K, thin).

4) Group Four: The instructor commented on the themes as in Group Three above. He conducted no theme discussion periods.

The three instructors are experienced, competent English teachers who adhered meticulously to the instructional methods prescribed for each group of students.

Each group of students consisted of one section of above-average students, one of average students and one of below-average students. Since the sectioning was determined by student scores on an objective literature test, there is no certain relation between a student's section and his potential or actual skill as a writer.

During the first semester, students wrote a total of ten themes. Six were written out of class and were 500-1000 words long; the other four were written in-class and were 300-400 words long. Each theme assignment was the same for all students and due at the same lesson.

At the end of the semester I asked the three instructors to identify students ineligible for participation in the project and discovered that some had come for extra help with their writing, some had resigned or been discharged, and one had lost his theme #1. The smallest section, after eliminating students who are ineligible, contains 10 students. Mr. Westen tells me that the interests of objective, reliable evaluation call for the other sections to be reduced to this size also. He will perform the reduction by a formal system of random selection.

Instructors graded their own themes and submitted the grades for record. But these grades are not a part of the project.

For purposes of this project, I will use the first, second, ninth and tenth themes (all four written out of class), 480 separate themes. These 480 will be reproduced on multilith mats and marked only with an identifying code. Then they will be graded by four graders who will have been trained in an intensive two-week course of instruction designed to prepare them to grade themes according to the components we use in teaching the students to write. The training course will first acquaint them with the grading criteria and then require them to grade themes, compare grades and discuss their grading agreements and differences until their results are satisfactorily close together.

Each grader will then be given the entire set of 480 themes and instructed to grade them at the rate of 75-100 themes per week. The themes will be arranged in haphazard sequence so that the graders will have no clue to the author or to the number of the theme.

The proposed time table is as follows:

August through December, 1964: The three instructors taught their twelve sections by the four different methods.

September, 1965: Typists will copy themes 1, 2, 9, and 10 on mats and make ten copies of each theme.

13 September 1965--24 September 1965: The four graders will attend a training course at the Air Force Academy to become familiar with the project and proficient at grading themes.

October, 1965--January, 1966: Graders will grade each of the 480 themes.

February--April, 1966: Mr. Westen will analyze the grade data.

May, 1966: Major Clark and Mr. Westen will prepare the final report and summary.

Mr. R. J. Westen, of the Academy Evaluation division, has prepared a comment on this proposal and an explanation of the several analyses he will make of data he receives from the graders. This comment-explanation follows:

> This experiment is designed to test the relative effectiveness of four methods of teaching English composition to Air Force Academy 4th Classmen (freshmen). The four methods, which have been discussed more extensively in preceding paragraphs, may be named for purposes of design exposition as: (1) Combined Method (involves both theme discussion and extensive written comments); (2) Theme Discussion Method (selected cadet themes discussed exhaustively in class); (3) Extensive Comment Method (all cadet themes receive extensive specific written instructor comments on strengths, weaknesses, and errors); and (4) Control Method (cadet themes receive only general written comments and marking of errors with symbols but are not discussed in class).
>
> The experiment will involve twelve sections of the Air Force Academy's regular 4th Class English course, with three sections being taught by each of the four methods. Subject will be assigned to the sections on an essentially random basis. (Actually, assignment will be made on the basis of scores on a test which in past years has proven uncorrelated with grades in this course.)
>
> Although each of the twelve sections will normally have 14 or 15 students, it is anticipated that usable data will be available on approximately 10 or 11 students per section. Part of the anticipated reduction in sample size is due to expected resignation or discharge of a few cadets. The remaining losses are expected to occur through absence from class (and consequent unavailability of data) of a few cadets and participation in extra instruction or tutoring sessions of a few others. Such participation would be expected to "contaminate" in an unknown fashion the teaching methods of the experiment and cadets receiving such instruction will be excluded. For ease in analyses of the data, each experimental section will be reduced by random selection of cases to be discarded to the N of the section

having the smallest number of usable cases. It is expected that the final experimental sample will contain 120 or 132 subjects.

Three instructors will handle all instruction of the experimental sections. Each instructor will be assigned to four of the sections and will instruct each of his sections by a different method. Determination of which section will be taught by which method will be made on a random basis.

Data for the experiment will be grades on four of the ten themes required during the course. Themes used in the study will be the first and second and the ninth and tenth. The first two themes, which will be written within the first five weeks of the course, will provide measures of proficiency before training. The ninth and tenth themes, to be written in the last three weeks of the course, will provide measures of proficiency after instruction. (The thirteen week period between the first and tenth themes will thus be the period of differential instruction for the four method groups. Themes 1, 9, and 10 are all of the same length, approximately 500 words. Each one requires the treatment of some relatively simple literary topic which stems from their study of a novel. Theme 2 is longer--700 to 1000 words--and calls for some research as well as the use of documentation and a bibliography.

The themes used in the experiment will be graded by the section instructors in their normal fashion for actual use in the course, but these grades will not be considered as criteria for the experiment. Rather, the themes will be reproduced with all student errors retained but with the instructors' markings omitted. The themes will be identified only by a code number which will permit identification by the experimenter of authors and sequence but which will not identify to anyone else either the authors or the sequence of the theme among the four assigned for the study. Copies of each theme will then be distributed to each of four readers who will mark them independently of each other.

The readers, who will receive prior training in the Academy's theme marking system, will mark each theme with three separate scores. Each score will be on a five-point scale with one as the lowest possible score and five as the highest score. The first score (I) will be an evaluation of purpose and organization, the second (II) an evaluation of content, and the third (III) an evaluation of sentences and diction. A total score (T), the sum of the other three scores, will also be recorded.

After all themes have been returned to the experimenter, they will be sorted by author and sequence of preparation and all scores collated, with each reader's scores being separately identified. Each theme will then have 16 scores, four from each of the four readers.

The first analyses will investigate the relationships among the sixteen scores for each of the four sets of themes. Agreement of readers will be determined separately for themes 1, 2, 9, and 10 for scores I, II, III, and T by computing the product-moment coefficients of correlation among all sixteen scores.

Next, separately for each reader, the independence of the three part scores he assigns over all themes will be checked. The last of these preliminary analyses will check the relationships over all subjects between scores on themes 1, 2, 9, and 10 for all readers.

Results of these preliminary analyses will greatly influence the subsequent analyses directly concerned with methods evaluation. Assuming satisfactory inter-reader agreement throughout all marking, the four readers' scores on a particular theme will be combined by addition so that each theme will now have only four scores, each of which is a total of marks given by the four readers.

If one of the four readers gives marks unrelated to marks given by the other three readers, the deviant reader's marks will be omitted. If the readers seem to split into pairs on the basis of their marking of themes, separate total scores for each pair of raters will be computed and used in subsequent analyses. Should there be little or no agreement between readers in their marking of themes, the experiment will be terminated at this point.

However, assuming that the pretraining of readers will be effective and result in substantial inter-reader agreement, the next point to be decided will be whether or not the three part-scores on each theme are sufficiently independent to merit separate analyses as criteria of English composition effectiveness. Should the correlations between any particular pair of part scores be as high as the agreement between readers for that pair of part scores, it could be argued that the part scores were not reliably different from each other and separate analyses of the particular pair of part scores would not be justified. It is thus possible that the criteria for each theme might be reduced from the four part-scores and the total score to a smaller number of variables, perhaps to the single total score.

The final decision to be made from the preliminary analyses will be whether or not scores on themes 1 and 2 and themes 9 and 10 should be treated separately or can be combined to give a single set of precourse measures and a single set of postcourse measures. If the coefficients of correlation between scores on themes 1 and 2 are higher than the correlations between scores on themes 1 and 9 or 2 and 10, it would indicate that differences in theme subject assignments have had a relatively minor effect and the precourse theme marks will be added across themes 1 and 2 to provide a single set of precourse measures. Similarly, marks will be added across themes 9 and 10 to provide a single set of postcourse measures.

It is obvious from the preceding section that the plans for the methods evaluation analyses must remain tentative until the results of the preliminary analyses can be evaluated. Assuming that these analyses justify each of them, the following actions will be taken:

(1) On each theme, the four readers' scores will be added together to give a total score for purpose and organization (T-I), a total score for content (T-II), a total score for sentences and diction (T-III), and a grand total score (T-T).

(2) For each cadet, analogous scores on themes 1 and 2 will be added together to give four precourse predictors.

(3) For each cadet, analogous scores on themes 9 and 10 will be added together to give four postcourse criteria.

(4) Four analyses of covariance will be run, one for each of the part scores and one for the total score. In each analysis of covariance, a precourse predictor will be used as a control variable in the analysis of the analogous criterion measure.

Although only minor random differences are expected between methods groups, the covariance technique adjusts criterion scores for any composition differences which may be initially present.

6. Personnel.

Director--Major William G. Clark, Course Director, English 111A and Associate Professor of English, USAF Academy, Colorado.

Associate Director--R. J. Westen, Chief of the Research Division, Evaluation Directorate, USAF Academy.

Graders--Four college graduates, preferably English majors, and preferably English teachers, either in high school or college. No effort has been made as yet to select graders. They will be selected on the basis of their understanding of English, conscientiousness, and willingness to adapt to what may be a radically different system of grading from what they have been using.

7. Facilities.

No facilities beyond those presently available at the Air Force Academy will be required.

8. Other Information.

a. No funds are known to be available for this project.

b. This proposal has not been submitted to any other agency or organization.

Appendix B
Results of Statistical Analyses
Project 5-8427

Correlational Analyses

Complete results of the correlational analyses are presented in Tables 1 through 5 of Appendix D. Table 1 summarizes the intra-reader and inter-reader agreement of the three part scores and the total score given by each reader to each of the 480 themes which comprised the experimental data; the table also shows the mean and standard deviation for each variable. Tables 2 through 5 present similar information for each of the four separate sets of themes. Each set consists of 120 themes written for one assignment.

To highlight the results of these analyses of intra-reader and inter-reader agreement of part and total scores, portions of Table 1 of Appendix D are presented below. Table A shows the degree of

TABLE A

CORRELATIONS OF PART SCORES
(ASSIGNED BY THE SAME READER ON ALL 480 THEMES*)

	Reader A			Reader B			Reader C			Reader D		
Parts	1	2	3	1	2	3	1	2	3	1	2	3
1	-	654	358	-	449	240	-	485	332	-	766	515
2		-	343		-	361		-	237		-	588

* Average intra-reader correlation among separate parts = .462.

correlation among the part scores assigned by each of the four readers. As an example, over the total population of 480 themes, Reader A's Part 1 (purpose and organization) scores correlated .654 with his Part 2 (content) scores; Reader A's Part 1 scores correlated .358 with his Part 3 scores (sentences and diction). Averaging these intra-reader agreement correlations by use of the

Notes: Decimal points omitted before all correlations in tables.

For N = 480, $r \geq .090$ is significant at the 5% level; $r \geq .118$ is significant at the 1% level.

z transformation, an average correlation of .462 was obtained. This statistically significant correlation indicates that, for a particular reader, themes given high scores on one part were normally given high scores on all three parts, while those given low scores on one part usually received low scores on all three parts. Reader D in particular showed this tendency to such an extent that one might question whether the three part scores were really assigned on the basis of the nominally separate characteristics which were to be evaluated or whether all part scores were strongly influenced by some one unspecified aspect of the theme.

TABLE B

CORRELATIONS AMONG CORRESPONDING PART SCORES
(ASSIGNED BY FOUR READERS ON ALL 480 THEMES*)

	Part Score 1				Part Score 2				Part Score 3			
Readers	A	B	C	D	A	B	C	D	A	B	C	D
A	-	275	375	401	-	353	497	387	-	239	423	342
B		-	133	334		-	202	325		-	152	178
C			-	328			-	298			-	207

* Average inter-reader correlation (reliability) between corresponding parts = .304.

Table B shows the extent of agreement between readers on the separate part scores assigned to the population of 480 themes. Thus, Reader A's Part 1 scores correlated .275 with Reader B's Part 1 scores, .375 with Reader C's Part 1 scores, and .401 with Reader D's Part 1 scores. Again using the z transformation in averaging correlations, the average inter-reader correlation of part scores was computed as .304. While there is statistically significant reliability of the part scores as indicated by these correlations showing agreement between readers, a reliability coefficient of .304 must be considered as relatively low.

The correlations shown in Tables A and B led to the conclusion that intra-reader agreement of the supposedly separate part scores was substantially greater than the reliability of the part scores so far as that reliability is measured by the inter-reader correlations. Since the average intra-reader agreement correlation (.462) was appreciably greater than the average inter-reader agreement correlation (.306), we cannot conclude that throughout the grading of the experimental themes the readers were continuing to maintain the separate identity of the three parts as they were established in the preliminary workshop. To go back to the specific example cited on the preceding page, had Reader A's Part 1 scores correlated more highly with Reader B's and Reader C's and Reader D's Part 1 scores

than they correlated with Reader A's Part 2 scores, we could have logically concluded that Part 1 (purpose and organization) constituted an aspect of theme composition which the readers were able to reliably identify, differentiate from other aspects of composition, and evaluate. Since this was not the case, we are not certain that the readers maintained common standards in the identification and evaluation of the separate aspects (or parts) of theme composition. For this reason it was decided to abandon use of the separate part scores and use only the total scores (sums of the three part scores) in subsequent analyses.

TABLE C

CORRELATIONS AMONG TOTAL SCORES
(ASSIGNED BY FOUR READERS ON ALL 480 THEMES*)

Readers	A	B	C	D
A	-	342	518	455
B		-	186	306
C			-	341

* Average inter-reader correlation (reliability) between total scores = .360.

Table C shows that there was slightly greater agreement between readers in terms of the total scores assigned to the population of 480 themes than there was between the separate part scores assigned by the same readers. The average inter-reader correlation of total scores was found to be .360. As reliability coefficients, these inter-reader agreement correlations all are statistically significant, but again they are lower than might have been hoped. However, by combining the total scores assigned by all four readers, one can arrive at a "grand total" score for each theme whose reliability can be estimated by the Spearman-Brown prophecy formula as equal to .620, which would normally be considered as acceptable reliability for an essay test.

Using only the one "grand total" score for each theme, correlations were computed between the four separate sets of themes written by the 120 subjects. These correlations are reported in Table D. As was noted on page 8 the assignments for Themes 1 and 5 were somewhat similar to each other while the assignments for Themes 2 and 6 were roughly parallel. Because of this assumed parallelism the original analysis plan contemplated using Theme 1 scores as an independent variable in an analysis of covariance of Theme 5 scores while Theme 2 scores were to be used as an independent variable in an analysis of co-variance of Theme 6 scores. In actual fact, criterion scores on Theme 1 were more closely related to scores on Theme 6 than to scores on Theme 5, while scores on

TABLE D

CORRELATIONS AMONG FOUR SETS OF THEME GRADES
(SCORES EOUAL SUMS OF TOTAL SCORES ASSIGNED BY ALL FOUR READERS)

Themes	1	2	5	6
1	-	499	218	331
2		-	314	279
5			-	342

Theme 2 were more closely related to scores on Theme 5 than to scores on Theme 6. Because of these results it was decided that scores on Themes 1 and 2 should be combined to provide a single pre-instruction variable while combining scores on Themes 5 and 6 would provide a single post-instruction variable. By pooling scores for each pair of themes, the reliability of each of the two final measures also increased. One method of approximating reliability, derived from correlations of inter-reader agreement on total scores for the separate sets of themes, yielded an estimated reliability of both the pre-instruction and the post-instruction variable slightly above .75.

The Analysis of Covariance

Table E presents results of the analysis of variance of the pre-instruction performance variable, the scores on Themes 1 and 2 combined. The F of 0.179 indicates the variability of scores within groups taught by the same method was appreciably greater than the variability among the different methods means. The hypothesis that the four groups do not differ significantly from one another in terms of means on the pre-instruction variable cannot be rejected.

TABLE E

ANALYSIS OF VARIANCE OF PRE-INSTRUCTION THEME SCORES*

Source of Variation	Sum of Squares	df	Mean Square	F
Between Methods	155	3	51.67	0.179
Within Methods	33,306	116	287.12	
Total	33,461	119		

* Scores are sums of Total Scores (all four readers' scores combined) on Themes 1 and 2 added together for each subject.

Table F presents results of the analysis of variance of the post-instruction criterion variable, the scores on Themes 5 and 6 combined. The F of 0.066 again indicates much more variability exists within groups taught by a particular method than exists among the means of the four methods groups. This finding again indicates the null hypothesis of no significant differences among methods means cannot be rejected.

TABLE F

ANALYSIS OF VARIANCE OF POST-INSTRUCTION THEME SCORES*

Source of Variation	Sum of Squares	df	Mean Square	F
Between Methods	44	3	14.67	0.066
Within Methods	25,620	116	220.86	
Total	25,664	119		

* Scores are sums of Total Scores (all four readers' scores combined) on Themes 5 and 6 added together for each subject.

The analysis of covariance might well have ended here, since a statistically significant F was highly improbable considering the results reported in the preceding two paragraphs. However, the analysis was carried out and the data are reported in Table G.

TABLE G

ANALYSIS OF COVARIANCE OF POST-INSTRUCTION THEME SCORES ADJUSTED FOR DIFFERENCES ON PRE-INSTRUCTION THEME SCORES

Source of Variation	Sums of Squares of Errors of Estimate	df	Mean Squares	F
Total	21,502	118		
Within Methods Groups	21,449	115	186.51	
Adjusted Methods Means	53	3	17.67	0.095

Correlation between Pre-Instruction Theme Scores and Post-Instruction Theme Scores = .403

TEST FOR HOMOGENEITY OF REGRESSION

Source of Variation	Sums of Squares	df	Mean Squares	F
Among Group Regressions	499.89	3	166.63	0.891
Deviations from Group Regression	20,949.11	112	187.04	

The F of 0.095 indicates that, allowing for initial differences in performance, there are still no statistically significant differences between means on the final criterion measure of the groups taught by the four different methods. The correlation between pre-instruction and post-instruction theme scores over all four methods groups is .403; the test for homogeneity of regression indicates that this correlation is relatively constant among the four groups.

Means and Standard Deviations

Since the variability in pre-instruction and post-instruction theme scores among all cadets taught by a particular method was so great, further study of the means and standard deviations of individual sections, of methods groups, and of groups taught by the same instructor was undertaken. Further, in an attempt to see if any initial differences in aptitude might have confounded the experimental findings, sections were grouped by "aptitude level," and means on the pre-instruction and post-instruction variables were computed. (As noted on page 3 , subjects were originally assigned to sections in the course on the basis of a test of knowledge of American and world literature whose scores in previous years' analyses had proved unrelated to theme grades but which provided homogeneous sections for the other aspects of the course dealing with literature.) Within the sections taught by each method, one section was a "high aptitude" section, one a "middle aptitude" section, and one a "low aptitude" section.

Mean and standard deviation data on the four separate themes appear only in Appendix D since the analysis of covariance involved combined theme scores. Table 6 of Appendix D presents means and standard deviations of total scores (totals of all part scores assigned by all four readers) for each section on each of the four themes. Means and standard deviations for methods groups are presented in Table 7 of Appendix D, while similar data for instructor groups (where all subjects taught by a particular instructor constitute an instructor group) are shown in Table 8 of Appendix D and means and standard deviations for "aptitude level" groups appear in Table 9 in the same appendix.

Table H presents means and standard deviations for each individual section and for instructional methods groups on the pre-instructional and post-instructional variables. Table I which follows presents means and standard deviations of the pre-instructional and post-instructional theme scores for groups of students taught by the separate instructors, while Table J presents similar data for students grouped by initial "aptitude level."

TABLE H

MEANS AND STANDARD DEVIATIONS
OF PRE-INSTRUCTION AND POST-INSTRUCTION THEME SCORES
FOR ALL SECTIONS GROUPED BY INSTRUCTIONAL METHOD

Method	Section Instructor	N	Pre-Instruction Scores (Themes 1 and 2) Mean	Pre-Instruction Scores (Themes 1 and 2) Std Dev	Post-Instruction Scores (Themes 5 and 6) Mean	Post-Instruction Scores (Themes 5 and 6) Std Dev
Combined	A	10	72.8	16.1	78.0	9.5
	B	10	73.0	16.7	80.8	8.8
	C	10	64.1	10.2	64.3	10.6
Total		30	70.0	15.2	74.4	12.1
Discussion	A	10	58.3	14.2	68.3	11.2
	B	10	83.4	14.8	82.6	14.2
	C	10	70.1	11.3	75.3	12.4
Total		30	70.6	17.0	75.4	13.9
Comment	A	10	65.4	14.8	66.7	13.3
	B	10	69.4	15.2	80.2	14.0
	C	10	69.9	21.1	79.8	12.4
Total		30	68.2	17.4	75.6	14.6
Control	A	10	60.1	18.6	65.7	16.4
	B	10	72.8	13.4	77.4	15.4
	C	10	70.8	15.6	79.6	16.7
Total		30	67.9	17.0	74.2	17.3
Total Sample		120	69.2	16.7	74.9	14.6

TABLE I

MEANS AND STANDARD DEVIATIONS
OF PRE-INSTRUCTION AND POST-INSTRUCTION THEME SCORES
FOR STUDENTS GROUPED BY INSTRUCTOR

Instructor	N	Pre-Instruction Scores (Themes 1 and 2) Mean	Pre-Instruction Scores (Themes 1 and 2) Std Dev	Post-Instruction Scores (Themes 5 and 6) Mean	Post-Instruction Scores (Themes 5 and 6) Std Dev
A	40	64.2	17.0	69.9	13.7
B	40	74.6	16.0	80.2	13.5
C	40	68.7	15.4	74.8	14.6

TABLE J

MEANS AND STANDARD DEVIATIONS
OF PRE-INSTRUCTION AND POST-INSTRUCTION THEME SCORES
FOR STUDENTS GROUPED BY "APTITUDE LEVEL"*

Aptitude Level	N	Pre-Instruction Scores (Themes 1 and 2) Mean	Pre-Instruction Scores (Themes 1 and 2) Std Dev	Post-Instruction Scores (Themes 5 and 6) Mean	Post-Instruction Scores (Themes 5 and 6) Std Dev
High	40	74.2	17.9	80.0	13.6
Middle	40	68.4	16.1	76.7	13.6
Low	40	64.9	14.6	68.0	14.0

* "Aptitude Level" measured by achievement on a locally-devised test of knowledge of American and world literature.

Appendix C
Discussion of Statistical Analyses
Project 5-8427

The Correlational Analyses

Although the readers were asked to grade separately on three aspects of theme writing, it was not assumed that these three aspects would be statistically independent. Rather, it was assumed that some positive correlation would exist among the three part scores, with Part 1 (purpose and organization) and Part 2 (content) expected to correlate more highly than Part 1 and Part 3 (sentences and diction) or Part 2 and Part 3 correlated. This expectation was confirmed by the analyses as was evidenced in Table 1 of the preceding section.

However, a priori logical definition of the separate parts had suggested that the parts were discriminably distinct. Use of the same part scores during reader training sessions had not revealed any problems in discriminating between the aspects to be graded. Thus, the experimental results indicating greater intra-reader agreement than inter-reader agreement came as a distinct surprise.

It is suggested that reader training sessions may not have been adequate to maintain the distinctions in the minds of the readers between the separate parts. Perhaps additional "standardization training sessions" introduced throughout the period when readers were grading the experimental themes would have helped maintain greater independence of parts. It should also be noted that the grading procedures introduced to reduce "halo effect" may not have been followed conscientiously by all readers or that the procedures themselves were inadequate to eliminate such effects.

The only measures of reliability used in the experiment were the measures of inter-reader agreement which were reported in the preceding section. However, the intra-reader agreement correlations also provide a basis for estimating reliability of an individual reader's part scores. Test theory demands that correlation between two score variables can be no higher than the product of their separate reliabilities. Thus, if test A has a reliability of .60 and test B has a reliability of .50, the correlation between A and B can be no greater than $.6 \times .5 = .30$, since random error variance of scores on one test will be uncorrelated with random error variance of scores on another test. The relatively high correlations among Reader D's part scores argue for reliabilities of part scores in the neighborhood of .75 to .85. However, such high reliability may simply indicate a systematic bias or non random error in scoring and be no proof of validity of the part scores. And, in view of the low agreement found between readers, the systematic bias explanation seems the

most reasonable interpretation for the high correlations among Reader D's part scores.

While the reliabilities of even total scores given by one reader over a single set of themes may appear quite low (.360), the increases in reliability obtained by pooling total scores of all four readers resulted in scores with what must be considered as quite satisfactory reliability for theme or essay grades. When scores for two sets of themes were pooled, the estimated reliability of approximately .75 is quite good for essay or theme grading albeit somewhat low when compared to reliabilities required of objective tests.

Reliabilities obtained in the present study are in line with those reported by Godshalk, Swineford, and Coffman (page 36) in a study which pooled judgments of a number of readers in evaluating compositions. The results of the present study taken together with those of the Godshalk study suggest that traditional methods of instructor evaluation of essays or compositions may well be inadequate; they at least suggest that instructors might exchange compositions among themselves so that each essay grade could be the result of pooled judgments of several instructors.

The Analysis of Covariance

The analysis of covariance was originally planned to take into account any differences between experimental groups in terms of ability and achievement in English composition prior to the experimental instruction. Although students were assigned to sections on the basis of a predictor variable which had proven uncorrelated with normal instructor grades in English composition, it was assumed that there might be, on a chance basis, significant differences between methods groups in terms of initial ability. The necessary randomness for analysis for covariance assumptions was generated by the fact that sections within mean levels of ability were assigned to methods groups on a random basis.

The analysis of covariance design did not provide for the evaluation of several interactions which, in fact, may have had some significant effect. These were the one three-way interaction of Instructor x Method x Aptitude Level and the two-way interactions of Instructor x Method, Instructor x Aptitude Level, and Method x Aptitude Level. Inspection of section means does not indicate that any of these interactions is likely to be a significant factor in the results obtained. However, the original design did not provide for statistical evaluation of the interaction effects and it cannot be concluded that these effects were not significant.

A basic assumption in analysis of variance is that within group variances are homogeneous from group to group. Although the tests of homogeneity of variance have not been reported in the results section, this assumption was tested prior to conducting each analysis of a variance and the null hypotheses concerning variances could not be rejected. The basic conditions for analysis of variance and covariance were satisfied.

Although one need not assume linearity of regression in carrying out an analysis of covariance, the assumption does become important in interpreting findings of an analysis of covariance. In the present study, the assumption of linearity could not be rejected, and we concluded that the relationship between pre-instruction grades and post-instruction grades was essentially the same for each of the four methods groups.

The only major conclusion of the entire analysis is that no significant differences between groups attributable to instructional methods were obtained. However, in view of the previous statement that the tests used to assign students to individual sections had proved to be unrelated to instructor composition grades, the finding of this study that sections differing in aptitude level as measured by this "sectioning" test did differ in theme grades assigned by the readers in the experiment and that "high aptitude" sections tended to get higher theme scores may be an ancillary finding of some practical significance. It is suggested that the positive correlation between "aptitude level" and theme performance obtained in this study is in part the result of standardization of theme grading and improved reliability of theme grades obtained by pooling scores of the four readers. It is suggested that normal theme grading by regular classroom instructors does not achieve nearly the same degree of standardization and consequent reliability. Absence of relationship between grades and the sectioning criterion is a necessary consequence of such lack of reliability and standardization.

Section Means and Individual Student Performance

As was noted in Appendix B, the investigators in the study went beyond the original design in order to try to obtain further insights into effects of the experimental methods. The analyses of variance and covariance indicated a high degree of variability within groups of students taught by the same methods compared to the observed variability between group means, and an effort was made to see what factors contributed to the large within group variance. The approach was simply to inspect carefully the pre-instruction score means and post-instruction score means of individual sections of methods groups, of instructor groups, and of aptitude level groups. In addition, the scores of individual students within sections were inspected carefully.

The following tentative conclusions, resulting from inspection of the scores and means are not conclusive but rather suggestive. One conclusion which has been previously cited is that aptitude level and both pre-instruction and post-instruction scores were significantly and positively correlated. Also previously mentioned was the conclusion that other interactions were probably not significant. A further finding from these inspections was that there were differences apparent in difficulty of the four themes assigned as part of the experiment, and though there may have been significant learning from pre-instruction measures to post-instruction measures, the last theme assigned (theme 6) was probably much more difficult for the students than theme 5. The explanation within the body of the text suggests that the additional difficulty of theme 6 was primarily a matter of a lack of structure in the assignment.

Examination of the scores of individual students rather than section and group means indicated extreme variability in performance of individual students over all four themes. In most educational experiments, students who are best on one criterion tend to be best on all measures. The same consistency in human performance was noticeably absent in the present study. Individuals performing extremely well on several themes often performed very poorly on the fourth, and these results suggest further problems in the experimental design.

The students whose themes were used in this experiment were not aware that they were participating in an experiment. The experiment itself extended over a prolonged period of approximately 16 weeks, and it must be assumed that subject motivation and aspiration level were significant factors which were entirely uncontrolled within the experimental design. While their effects were probably random with respect to evaluation of methods differences, they operated to produce the extremely high within group variability. It is suggested that there were probably major differences from subject to subject and even within subjects in terms of the amount of effort expended in the preparation of the themes used in the experiment and that this effort was a function of other environmental demands upon the experimental subjects and the subjects' aspirational levels. Some students assured of receiving satisfactory grades in the course at the time of preparation of the last theme may well have let down in their efforts in the preparation of this theme, while other subjects probably felt it necessary to do well on this last theme in order to pass the course or to improve their grades.

The experimental design, in attempting to utilize the normal educational procedures and environment, restricted experimental control of subject motivation and effort, and it is suggested that the factor of motivation was at least as important as the factor of instructional method in producing the results which were obtained.

$r \geq .118$ is significant at the 1% level.

Appendix D
Statistical Tables
Project 5-8427

highly with Reader B's and Reader C's and Reader D's Part 1 scores

TABLE 1

PRODUCT MOMENT CORRELATIONS* AMONG SUBSCORES AND TOTAL SCORES ASSIGNED BY FOUR RATERS OVER 480 THEMES

Variable	Variable Number	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Mean	Standard Deviation
Reader A - Part 1	1	654	358	842	275	308	120	309	375	355	125	378	401	392	244	389	2.99	1.39
Part 2	2	-	343	835	233	353	107	304	384	497	129	446	342	387	222	355	2.97	1.39
Part 3	3		-	712	088	134	239	204	294	216	423	413	261	299	342	346	3.00	1.39
Total	4			-	249	333	196	342	441	446	284	518	420	451	338	455	8.96	3.32
Reader B - Part 1	5				-	449	240	748	133	142	-021	112	334	310	139	290	2.97	1.40
Part 2	6					-	361	799	240	202	002	196	259	325	136	266	3.00	1.38
Part 3	7						-	711	045	051	152	110	087	107	178	134	3.12	1.40
Total	8							-	186	174	061	186	301	328	203	306	9.08	3.14
Reader C - Part 1	9								-	485	332	804	328	314	177	309	3.02	1.41
Part 2	10									-	237	761	273	298	166	288	3.02	1.40
Part 3	11										-	695	116	129	207	174	2.99	1.41
Total	12											-	317	327	243	341	9.03	3.18
Reader D - Part 1	13												-	766	515	875	2.99	1.42
Part 2	14													-	588	894	2.99	1.41
Part 3	15														-	798	2.99	1.42
Total	16															-	8.95	3.69

* Decimal points omitted before all correlation coefficients.
For N=480, $r \geq .090$ is significant at the 5% level and $r \geq .118$ is significant at the 1% level.

TABLE 2

PRODUCT MOMENT CORRELATIONS* AMONG SUBSCORES AND TOTAL SCORES ASSIGNED BY FOUR RATERS OVER 120 THEMES ON ASSIGNMENT 1

Variable	Variable Number	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Mean	Standar Deviati
Reader A - Part 1	1	607	359	829	254	184	218	291	274	345	217	358	438	457	305	478	2.75	1.37
Part 2	2	-	340	811	246	370	204	361	279	436	126	358	355	382	192	368	2.91	1.30
Part 3	3		-	729	092	270	331	307	395	313	453	502	213	314	423	386	2.77	1.41
Total	4			-	248	346	320	405	402	460	342	517	423	487	393	521	8.42	3.22
Reader B - Part 1	5				-	454	163	714	229	170	076	205	471	416	221	441	2.89	1.41
Part 2	6					-	437	832	164	360	101	264	270	304	169	295	2.55	1.38
Part 3	7						-	714	173	143	203	224	099	181	154	173	3.04	1.44
Total	8							-	251	296	169	307	371	398	241	402	8.48	3.18
Reader C - Part 1	9								-	480	453	842	270	257	206	294	3.02	1.48
Part 2	10									-	258	734	227	319	138	269	3.48	1.33
Part 3	11										-	742	129	250	268	261	3.08	1.42
Total	12											-	270	354	265	355	9.58	3.29
Reader D - Part 1	13												-	675	400	829	2.35	1.31
Part 2	14													-	528	871	2.32	1.18
Part 3	15														-	791	2.68	1.39
Total	16															-	7.35	3.21

* Decimal points omitted before all correlation coefficients.
For N=120, $r \geq .180$ is significant at the 5% level and $r \geq .236$ is significant at the 1% level.

TABLE 3

PRODUCT MOMENT CORRELATIONS* AMONG SUBSCORES AND TOTAL SCORES ASSIGNED BY FOUR RATERS OVER 120 THEMES ON ASSIGNMENT 2

Variable	Variable Number	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Mean	Standard Deviation
Reader A - Part 1	1	691	255	833	319	439	082	341	491	444	-019	425	475	535	252	440	3.01	1.43
Part 2	2	-	288	855	299	401	183	360	359	620	-008	447	401	489	307	416	2.88	1.52
Part 3	3		-	649	004	152	162	127	269	286	471	481	195	235	325	281	2.69	1.41
Total	4			-	266	427	185	356	479	580	186	579	457	538	375	485	8.59	3.40
Reader B - Part 1	5				-	544	278	781	130	194	-098	103	251	322	061	201	3.00	1.40
Part 2	6					-	506	843	291	228	-022	231	321	386	260	334	3.58	1.16
Part 3	7						-	749	099	040	127	126	104	147	193	127	3.06	1.29
Total	8							-	217	188	007	192	281	357	216	276	9.62	3.05
Reader C - Part 1	9								-	492	196	794	386	434	284	404	3.20	1.42
Part 2	10									-	097	731	457	507	325	504	2.68	1.33
Part 3	11										-	613	029	110	334	174	2.63	1.41
Total	12											-	404	488	441	503	8.51	2.96
Reader D - Part 1	13												-	785	518	871	2.94	1.34
Part 2	14													-	573	855	2.92	1.40
Part 3	15														-	753	2.85	1.38
Total	16															-	8.62	3.61

* Decimal points omitted before all correlation coefficients.
For N=120, $r \geq .180$ is significant at the 5% level and $r \geq .236$ is significant at the 1% level.

TABLE 4

PRODUCT MOMENT CORRELATIONS* AMONG SUBSCORES AND TOTAL SCORES ASSIGNED BY FOUR RATERS OVER 120 THEMES ON ASSIGNMENT 5

Variable	Variable Number	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Mean	Standard Deviation
Reader A - Part 1	1	557	318	810	273	365	126	337	366	365	019	326	313	244	161	283	3.30	1.36
Part 2	2	-	339	813	097	392	135	273	451	507	191	493	292	370	253	362	3.29	1.32
Part 3	3		-	706	040	163	237	195	322	141	342	348	242	255	205	278	3.33	1.31
Total	4			-	178	396	213	346	480	436	234	500	364	372	265	396	9.92	3.09
Reader B - Part 1	5				-	418	260	749	-020	072	-110	-024	279	272	168	284	3.20	1.45
Part 2	6					-	356	778	258	215	-058	179	170	310	114	233	3.31	1.37
Part 3	7						-	723	068	075	061	089	181	191	251	249	3.29	1.45
Total	8							-	132	159	-047	106	281	343	238	341	9.80	3.20
Reader C - Part 1	9								-	483	364	793	353	250	074	264	3.19	1.30
Part 2	10									-	292	782	209	228	129	223	3.22	1.43
Part 3	11										-	722	080	-002	073	061	3.22	1.38
Total	12											-	276	206	121	238	9.62	3.15
Reader D - Part 1	13												-	718	457	857	3.72	1.30
Part 2	14													-	497	872	3.71	1.28
Part 3	15														-	790	3.42	1.41
Total	16															-	10.86	3.35

* Decimal points omitted before all correlation coefficients.
For N=120, $r \geq .180$ is significant at the 5% level and $r \geq .236$ is significant at the 1% level.

TABLE 5

PRODUCT MOMENT CORRELATIONS* AMONG SUBSCORES AND TOTAL SCORES ASSIGNED BY FOUR RATERS OVER 120 THEMES ON ASSIGNMENT 6

Variable	Variable Number	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Mean	Standard Deviation
Reader A - Part 1	1	754	486	890	211	205	018	195	354	337	282	424	290	250	184	274	2.88	1.37
Part 2	2	-	404	857	242	269	-122	171	460	431	191	471	261	268	068	227	2.78	1.57
Part 3	3			752	214	026	196	205	261	170	381	355	269	274	325	328	3.22	1.54
Total	4			-	267	201	035	227	451	376	340	500	328	317	229	331	8.89	3.40
Reader B - Part 1	5				-	402	248	750	137	151	036	132	310	222	053	222	2.79	1.31
Part 2	6					-	219	735	159	194	091	194	131	215	-035	119	2.54	1.30
Part 3	7						-	692	-185	-066	200	-021	-100	-129	076	-059	3.08	1.40
Total	8							-	044	114	153	136	150	134	045	125	8.42	2.91
Reader C - Part 1	9								-	535	364	827	334	338	136	307	2.68	1.36
Part 2	10									-	233	769	370	333	122	313	2.71	1.36
Part 3	11										-	698	194	140	117	170	3.02	1.37
Total	12											-	391	353	164	344	8.42	3.13
Reader D - Part 1	13												-	751	571	878	2.93	1.39
Part 2	14													-	663	916	3.00	1.44
Part 3	15														-	846	3.02	1.41
Total	16															-	8.96	3.74

* Decimal points omitted before all correlation coefficients.
For N=120, $r \geq .180$ is significant at the 5% level and $r \geq .236$ is significant at the 1% level.

TABLE 6

MEANS AND STANDARD DEVIATIONS OF TOTAL SCORES (ALL READERS' SCORES COMBINED) FOR INDIVIDUAL SECTIONS ON EACH OF FOUR ASSIGNED THEMES

Method	Section Instructor	Theme 1 Mean	Theme 1 Std Dev	Theme 2 Mean	Theme 2 Std Dev	Theme 5 Mean	Theme 5 Std Dev	Theme 6 Mean	Theme 6 Std Dev
Combined	A	36.0	9.2	36.8	9.1	40.0	6.6	38.0	8.0
	B	36.1	10.2	36.9	11.1	44.9	7.3	35.9	6.8
	C	28.9	3.8	35.2	7.8	32.8	6.5	31.5	6.9
Discussion	A	27.8	6.9	30.5	8.7	37.1	5.9	31.2	7.7
	B	40.6	8.4	42.8	7.9	43.0	9.5	39.6	8.3
	C	33.3	8.6	36.8	4.5	41.5	6.0	33.8	8.1
Comment	A	32.4	8.2	33.0	8.8	34.6	10.4	32.1	5.9
	B	35.5	10.9	33.9	6.5	44.2	8.5	36.0	9.2
	C	35.4	11.6	34.5	12.8	42.5	6.7	37.3	8.9
Control	A	29.8	10.2	30.3	11.0	35.1	9.4	30.6	8.6
	B	36.7	7.4	36.1	9.6	44.1	5.9	33.3	10.9
	C	33.5	9.7	37.3	8.6	42.7	10.5	36.9	10.8
Total Sample		33.8	9.6	35.3	9.6	40.2	8.9	34.7	8.9

TABLE 7

MEANS AND STANDARD DEVIATIONS OF TOTAL THEME SCORES
OF GROUPS TAUGHT BY DIFFERENT METHODS ON FOUR ASSIGNED THEMES

		Theme 1		Theme 2		Theme 5		Theme 6	
Method	N	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Combined	30	33.7	8.9	36.3	9.4	39.2	8.4	35.1	7.7
Discussion	30	33.9	9.5	36.7	8.8	40.5	7.7	34.9	8.8
Comment	30	34.4	10.4	33.8	9.7	40.4	9.6	35.1	8.4
Control	30	33.3	9.6	34.6	10.2	40.6	9.6	33.6	10.5
Total Sample	120	33.8	9.6	35.3	9.6	40.2	8.9	34.7	8.9

TABLE 8

MEANS AND STANDARD DEVIATIONS OF TOTAL THEME SCORES
OF GROUPS TAUGHT BY DIFFERENT INSTRUCTORS ON FOUR ASSIGNED THEMES

		Theme 1		Theme 2		Theme 5		Theme 6	
Instructor	N	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
A	40	31.5	9.2	32.6	9.8	36.7	8.6	33.0	8.2
B	40	37.2	9.5	37.4	9.5	44.0	7.9	36.2	9.2
C	40	32.8	9.2	36.0	9.0	39.9	8.7	34.9	9.1

TABLE 9

MEANS AND STANDARD DEVIATIONS OF TOTAL THEME SCORES
OF STUDENTS GROUPED BY "APTITUDE LEVEL"* ON FOUR ASSIGNED THEMES

		Theme 1		Theme 2		Theme 5		Theme 6	
Aptitude Level	N	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
High	40	36.4	10.1	37.8	10.2	42.0	8.6	38.0	9.1
Middle	40	34.0	9.7	34.4	9.4	42.6	7.7	34.1	9.0
Low	40	31.1	8.2	33.8	8.7	36.0	8.9	32.0	7.5

* "Aptitude Level" measured by achievement on a locally-devised test of knowledge of American and world literature.